Structure
Ways & Means

Ab Initio Structure Determination from Electron Microscopic Images of Single Molecules Coexisting in Different Functional States

Dominika Elmlund,[2,4] Ralph Davis,[1] and Hans Elmlund[1,3,*]
[1]Department of Structural Biology, Fairchild Building, Stanford University School of Medicine, Stanford, CA 94305, USA
[2]Department of Laboratory Medicine, Karolinska Institutet, Novum, SE-141 86 Huddinge, Sweden
[3]Department of Medical Biochemistry and Cell Biology, Göteborg University, P.O. Box 440, SE-405 30 Göteborg, Sweden
[4]Department of Biosciences and Nutrition, Karolinska Institutet and School of Technology and Health, Royal Institute of Technology, Novum, SE-141 87 Huddinge, Sweden
*Correspondence: hael@stanford.edu
DOI 10.1016/j.str.2010.06.001

SUMMARY

We have developed methods for ab initio three-dimensional (3D) structure determination from projection images of randomly oriented single molecules coexisting in multiple functional states, to aid the study of complex samples of macromolecules and nanoparticles by electron microscopy (EM). New algorithms for the determination of relative 3D orientations and conformational state assignment of single-molecule projection images are combined with well-established techniques for alignment and statistical image analysis. We describe how the methodology arrives at homogeneous groups of images aligned in 3D and discuss application to experimental EM data sets of the *Escherichia coli* ribosome and yeast RNA polymerase II.

INTRODUCTION

Three-dimensional (3D) structure determination of large biological assemblies by electron microscopy (Crowther et al., 1970b; DeRosier and Klug, 1968; Frank, 2006; Hoppe et al., 1976; van Heel et al., 2000) is plagued by the problem of establishing relative 3D orientations and assigning conformational state to many thousands of noisy single-molecule projection images. Imaging by cryo-electron microscopy (cryo-EM) protects the particles from deformation and excessive radiation damage by embedment in vitreous ice (Adrian et al., 1984; Knapek and Dubochet, 1980). In contrast to crystallography, where the crystal restricts conformational variability, cryo-EM images represent the structural variability inherent in the solution state. With access to the scattering information from each individual molecule, methods can be developed to define subpopulations of molecules with distinct structural characteristics. The characterization of structural variations will be necessary both for a full mechanistic understanding and for achieving near-atomic resolution cryo-EM reconstructions of many biological macromolecules.

The successful assignment of conformational state to series of 2D macromolecular projections depends on the quality of the 3D alignment. To date, many computational methods for heterogeneity analysis of asymmetric molecules in cryo-EM have relied heavily on 3D reference volumes for providing the initial 3D alignment (Brink et al., 2004; Fu et al., 2007; Gao et al., 2004; Hall et al., 2007; Penczek et al., 2006; Scheres et al., 2007; Shatsky et al., 2010; Zhang et al., 2008). Classification of synthetically generated heterogeneous projections with known rotational centers into homogeneous subsets has been accomplished without using 3D reference volumes (Herman and Kalinowski, 2007). In practice, however, for N particle images all $6N$ degrees of freedom need to be considered simultaneously, the angular orientations, the origin shifts, and the conformational state assignments. In the special case of heterogeneity that requires separation of images of factor-bound complexes from images of empty complexes, homogeneous subsets of images may be identified by 2D multivariate statistical analysis, before 3D reconstruction (Elad et al., 2008).

We strived to overcome the need for a priori structural or heterogeneity information in single-particle reconstruction from EM data of asymmetric molecules and designed a methodology that attempts to avoid the model bias that is inherent to all template-based alignment techniques. In our procedure, well-established techniques for projection matching (Penczek et al., 1994) and statistical image analysis (Frank and van Heel, 1982; Lebart et al., 1984; van Heel and Frank, 1981) are combined with improved algorithms for reference-free 3D alignment and conformational state assignment. Our algorithms rely on the "projection slice theorem" (Bracewell, 1956), which states that any two nonparallel 2D projections of the same 3D object will share a common line in Fourier space. In the absence of a suitable reference volume, a common approach is the reference-free 2D alignment (Penczek et al., 1992) and classification of particles into classes representing common projection directions (Frank and van Heel, 1982; Lebart et al., 1984; van Heel and Frank, 1981). Unsupervised classification procedures of this kind allows for improvement of the signal-to-noise ratio by averaging. The first ab initio method for determination of the 3D orientations of class averages of asymmetric structures was the method of angular reconstitution (van Heel, 1987), which establishes the relative angular orientations of three projections via common lines-based orientation search. Approximation algorithms enable inclusion of a large number of class averages in a single reference-free 3D alignment step (Elmlund et al., 2008; Ogura and Sato, 2006; Penczek et al., 1996; Singer et al., 2009).

The advantages of these recent approaches over finding an explicit solution for only a few averages include increasing the effective signal-to-noise ratio of the 3D alignment, and reducing the risk of computing an erroneous alignment optimum because of an unfortuitous selection of class averages.

In the present study, we use a Fourier space formulation of the projection slice theorem to enable coupling of a parameter that controls the direction of the angular orientation search with a low-pass filter that improves the robustness of the reference-free 3D alignment toward noise. Random sampling of the projections included for calculation of the common line correlation coefficient improves the computational efficiency. Restriction of the number of class averages that are allowed to be assigned the same projection direction accounts for the geometrical constraint that a class average normally represents a distinct view or conformational state. A 3D alignment is established in a discrete angular space before moving to a *6K* dimensional search space to maximize the spectral common line correlation coefficient over *5K* continuous orientation parameters (*3K* Euler angles and *2K* translational degrees of freedom) and *K* discrete state assignment parameters, where *K* is the number of class averages. The method is applied to one synthetic (Baxter et al., 2009) and two experimental EM data sets: cryo-EM data of the *Escherichia coli* 70S ribosome bound to elongation factor G (Gao et al., 2004) and negative stain data of yeast RNA polymerase II in its 12-subunit form, including the dissociable Rpb4/7 module (Armache et al., 2005; Bushnell and Kornberg, 2003; Edwards et al., 1991). With the synthetic data set, we find a distribution of particle images that is in perfect agreement with theory (99.7% correctly assigned images), and we produce ab initio ribosome reconstructions from experimental cryo-EM data that are consistent with previously published results (Elad et al., 2008; Penczek et al., 2006). Finally, we reconstruct two distinct conformations of 12-subunit yeast RNA polymerase II, one of which has not been previously observed.

METHODS

Reference-free 3D Alignment of Class Averages in a Discrete Angular Space

The simulated annealing (SA) optimization algorithm on which RAD is based has been described previously (Elmlund et al., 2008). In brief, the SA optimizer in RAD evolves from high ($T = 1{,}000{,}000$) to low ($T = 0.01$) temperature (T), using an exponential cooling schedule (Kirkpatrick et al., 1983):

$$T_{it+1} = 0.9T_{it}. \tag{1}$$

where *it* is the iteration counter. Random variables are used to assign angular orientations (Metropolis and Ulam, 1949). A constant number of rearrangements are performed at each temperature level. All changes which lead to improved common line correlation are accepted. If a trial solution is worse than the current one the transition is accepted with a temperature-dependent probability *P*, according to the Metropolis rule (Metropolis et al., 1953):

$$P = \min\left(1, \exp\left\{\frac{(C_{it+1} - C_{it})}{kT}\right\}\right), \tag{2}$$

where *C* is the common line correlation coefficient (defined in Equation 4, below), $k = 10^{-4}$ is the Bolzmann factor. The transition mechanism of RAD, in which a single class average changes its angular orientation, is redefined to account for the geometrical constraint that a class average normally represents a distinct view or conformational state. Let *K* be the number of class averages and let *P* be the set of *2K* projection directions in a spiral-like distribution (Saff and Kuijlaars, 1997). Let *S* be a subset of *P* ($S \subset P$), consisting of *K* mutually distinct projection directions, each participating in the current solution. A projection direction is then selected randomly for addition to the current solution from the $P - S = \{x \in P | x \notin S\}$ set. If the number of conformational states *t* is larger than one, *P* is modified to contain each direction in *t* copies, while preserving its size. This limits the number of class averages that can be assigned the same projection direction in RAD to the number of conformational states tested. As the number of states increases, the probability that the search will test assignment of class averages to the same projection direction increases proportionally. To understand the effect of this constraint, consider the simplified general problem of labeling *K* objects using *2K* possibilities for each object, giving $(2K)^K$ possible solutions. Introducing the constraint that no two labels are allowed to be equal reduces the number of possible solutions *F* to:

$$F = \prod_{f=K+1}^{2K} f. \tag{3}$$

The marked improvement of solution quality when aligning large populations of class averages using this constraint is explained by the reduction of the search space size. The reduction would, however, not be possible to apply to the alignment of individual particle images without prior classification. We initialize the one-exchange neighborhood SA algorithm with a random solution, and use the spectral correlation, calculated over common lines between the projection with altered angular orientation *p* and the remaining, "fixed" projections, for evaluating a transition:

$$C_{SA} = \frac{\sum_{\substack{i=1 \\ i \neq p}}^{K} \sum_{m=r^{hp}}^{r_T^{lp}} \mathrm{Re}\left\{c_{i,m,o_i} c'^{*}_{p_i,m,o_{p_i}}\right\}}{\sqrt{\sum_{\substack{i=1 \\ i \neq p}}^{K} \sum_{m=r^{hp}}^{r_T^{lp}} |c_{i,m,o_i}|^2 \sum_{\substack{i=1 \\ i \neq p}}^{K} \sum_{m=r^{hp}}^{r_T^{lp}} |c'^{*}_{p_i,m,o_{p_i}}|^2}} \tag{4}$$

where $o_i = \{\psi_i, \theta_i, \varphi_i\}$ *and* $0 \leq \varphi_i, \psi_i \leq 2\pi$ $0 \leq \theta_i \leq \pi$ *i* denotes the index of the Fourier transformed class average, $p_i \in \{1, 2, \ldots, K\}$ is the index of the Fourier transformed class average with altered angular orientation, *m* is the Fourier index, ψ, θ, φ are the three Euler angles, and

$$\begin{aligned} c_{i,m,o_i} &= A_{i,m,o_i} + i\Phi_{i,m,o_i} \\ c'_{p_i,m,o_{p_i}} &= A'_{pi,m,o_{p_i}} + i\Phi'_{pi,m,o_{p_i}} \end{aligned} \tag{5}$$

describe the real *A* and complex Φ parts of the common line components in the respective coordinate systems of the planes *i* and *p* at Fourier index *m*. r^{hp} and r_T^{lp} are the high-pass and low-pass Fourier index limits, respectively. The temperature-dependence of the low-pass Fourier index limit is discussed below. The algebra used to generate common lines has been described

elsewhere (Elmlund and Elmlund, 2009). The sampling points do not in general coincide with the sampling points of the Fourier transforms of the projections, and the sinc-function is used as a convolution kernel to retrieve common line Fourier components. Application of the Nyquist-Shannon sampling theorem in common lines-based orientation search has been described by others (Crowther et al., 1970a; Lindahl, 2001).

To reduce the computational burden of aligning a large population of class averages, we implemented a temperature-dependent random sampling approach. In this approach, the class average with altered orientation p is aligned to a randomly selected subset of the "fixed" class averages. The sample size (SZ_T) is linked to the temperature state of the annealing, so that few projections are used in scoring at high temperature, where the acceptance probability of any transition is high, but more projections are used at lower temperature, when a more accurate estimate of the scoring function is required. The random sample is updated at each change of temperature using the following scheme for calculating the sample size:

$$SZ_T = \begin{cases} \min\left(100, \frac{100}{\log(T)}\right) & \text{if } T > 1.0 \\ 100 & \text{if } T \leq 1.0 \end{cases}. \quad (6)$$

The resolution window used for calculation of the spectral common line correlation in RAD is coupled to the temperature control parameter. The high-pass frequency limit neither critically influences the outcome of an alignment, when kept in an interval between the diameter of the particle and up to about 80 Å, nor affects the computation time significantly. Thus, it can be fixed at a constant value throughout the alignment. In contrast, the low-pass frequency limit is an important variable, both for the stable convergence of RAD, and for the computational effort required. A temperature-coupled dynamic low-pass filter should exclude high frequencies at high temperatures and include Fourier components of higher frequency in the low-temperature transitions. We therefore propose the following temperature-dependent update scheme for the low-pass frequency limit (r_T^{lp}):

$$r_T^{lp} = \begin{cases} r_{\min}^{lp} + \frac{\log(T)}{\log(T_{init})}\left(\frac{r^{hp}}{2} - r_{\min}^{lp}\right) & \text{if } T > 1.0 \\ r_{\min}^{lp} & \text{if } T \leq 1.0 \end{cases} \quad (7)$$

where $r_{\min}^{lp}$ and r^{hp} are the user-defined low-pass and high-pass frequency limits, respectively, with typical values being $r_{\min}^{lp} = 20.0$A and $r^{hp} = 120.0$A. T is the current temperature ($1{,}000{,}000 \geq T \geq 0.01$) and T_{init} is the initial temperature. The search parameters are plotted together with temperature, average transition probability, and correlation as functions of iteration in Figure S1.

Maximization of a Multidimensional Common Line Correlation Coefficient

To date, the common lines-based assignment of conformational state to experimental cryo-EM projections of asymmetric molecules has relied on 3D alignment to a single reference volume generated from the heterogeneous data (Elmlund et al., 2009; Hall et al., 2007; Lundqvist et al., 2010; Shatsky et al., 2010). In contrast, the approach presented here does not require calculation of volumes from heterogeneous data, since conformational state assignment is achieved by solving the following multidimensional optimization problem:

$$o^* = \max_o C_{GRASP/DE}(o) = \max_o \frac{\sum_{i=1}^{K-1}\sum_{\substack{j=i+1 \\ \text{if } s_i = s_j}}^{K}\sum_{m=r^{hp}}^{r_{\min}^{lp}} \mathrm{Re}\left\{c_{i,m,o_i} c_{j,m,o_j}^{r^*}\right\}}{\sqrt{\sum_{i=1}^{K-1}\sum_{\substack{j=i+1 \\ \text{if } s_i = s_j}}^{K}\sum_{m=r^{hp}}^{r_{\min}^{lp}} |c_{i,m,o_i}|^2 \sum_{i=1}^{K-1}\sum_{\substack{j=i+1 \\ \text{if } s_i = s_j}}^{K}\sum_{m=r^{hp}}^{r_{\min}^{lp}} |c_{j,m,o_j}^{r^*}|^2}} \quad (8)$$

where $o = \{o_i\}_{i=1}^{K}$ and $o_i = \{\psi_i, \theta_i, \varphi_i, x_i, y_i, s_i\}$

$$0 \leq \varphi_i, \psi_i \leq 2\pi \quad 0 \leq \theta_i \leq \pi \quad -sh_{\max} \leq x_i, y_i \leq sh_{\max}.$$

A $6K$-dimensional spectral common line correlation coefficient is maximized to determine $5K$ continuous orientation degrees of freedom and K discrete state assignment parameters, where K is the number of class averages. ψ, θ, φ are the three Euler angles, and $sh_{\max} \in \mathbb{R}$ is an externally defined constant that controls the degree of translation. S_i is the integer state assignment parameter. The conditional statement *if* $s_i = s_j$ in the second summation over j ensures that only common lines between projections assigned to the same state are contributing to the correlation coefficient. The maximization begins with obtaining a state assignment solution by greedy randomized adaptive local search (GRASP) (Feo and Resende, 1995) in a one-exchange neighborhood, while keeping the angular orientations obtained by RAD fixed and assuming zero origin shifts. Each GRASP iteration consists of constructing a trial solution and then applying an exchange procedure to find a local optimum (i.e., the final solution for that iteration). A first random solution is generated and then solution element exchanges are considered in randomized order. Each element exchange must give an improved solution as measured by the common line correlation. The process is iterated until no improving moves can be found for a given number of iterations. The search is adaptive because the element chosen for addition to the current solution at any iteration is a function of those previously chosen. The conformational assignment solution is next inputted together with the discrete RAD alignment to a similar GRASP scheme, equipped with a differential evolution (DE)-based optimizer, operating over a projection's six degrees of freedom. The DE-optimizer is an extension of our previously published algorithm (Elmlund and Elmlund, 2009) that has been modified to account for the integer state parameter by a trivial continuous to discrete mapping (see Supplemental Information available online). Each GRASP iteration now constructs a trial solution using the DE optimizer. The converged solution is used to calculate volumes of the different conformational states

by back-projection (Harauz and van Heel, 1986; Radermacher, 1992).

Refinement of the In-Plane Alignment Parameters by Projection Matching

If high-quality in-plane parameters have been established, large data sets of electron microscopic projections can be efficiently grouped into homogeneous classes in 2D by using standard techniques for statistical image analysis (Frank and van Heel, 1982; Lebart et al., 1984; van Heel and Frank, 1981). This is true when the variations of the data set are due to difference in *both* projection direction and conformational state. The reference-free 2D alignment algorithm developed by Penczek et al. (1992) can be used to provide in-plane parameters of sufficient quality for classification to resolve heterogeneity due to differences in projection direction. Any template generated from these first class averages will, however, be biased by conformational heterogeneity if present, and the question arises whether such a biased reconstruction can be used to for refinement of the in-plane parameters via projection matching. An iterative scheme is proposed, where the above described methods are combined with projection matching (Penczek et al., 1994) onto the ab initio generated templates in order to refine the in-plane parameters *only*, and progressively improve class homogeneity, see Figure 1 for a flowchart of the method. In the present study, we show that the combination of RAD and GRASP/DE-based state assignment with projection matching enables ab initio reconstruction of the different conformational states present in the population.

Implementation of the concepts described above was done in SIMPLE (Single-particle IMage Processing Linux Engine), which is an object-oriented Fortran 95/2003 library that is freely available upon request. Shell scripts controlling the distribution of jobs for different state groups were executed in parallel on the LUNARC Linux cluster at Lund University. Seven hundred twenty CPU hours were used for reconstruction of pol II, while testing for five state groups in each iteration.

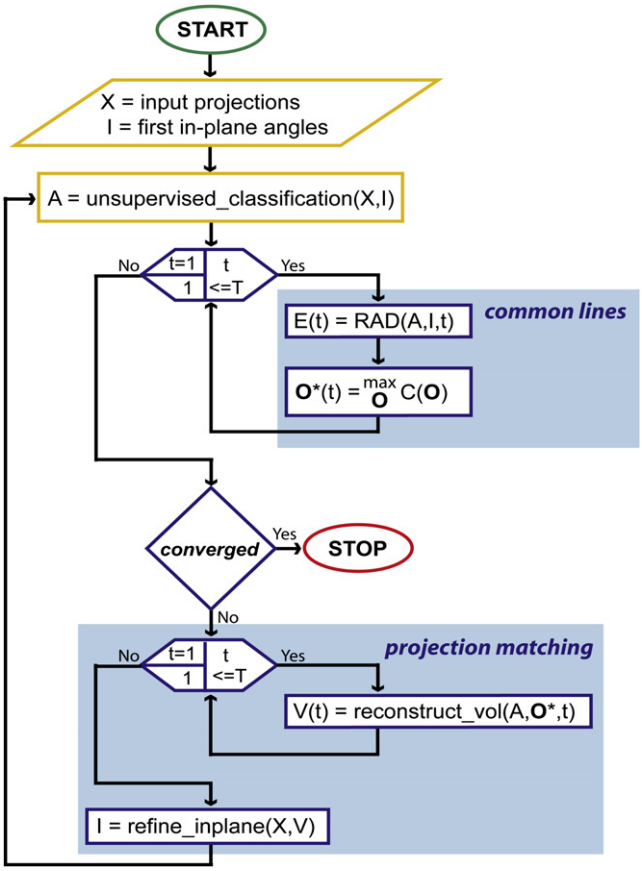

Figure 1. Flowchart for the Method

In the rhombic box, initialization of variables is performed, rectangular boxes describe computations, and diamond boxes represent decision nodes. The hexagon divided into three regions represent a loop, with the initialization defined in the upper left region, the increment defined in the lower left region, and the stopping condition defined in the rightmost region. *X* is the set of input projections and *I* is their assigned in-plane parameters. *A* is the set of class averages generated by unsupervised classification. *t* is the number of states to resolve and *T* is the upper limit of *t*. For each *t* a 3D alignment *E(t)* is obtained by RAD and the discrete solution is fed to a program maximizing a *6K*-dimensional spectral common line correlation coefficient *C(**O**)* (Equation 8) that determines *5K* continuous orientation degrees of freedom and *K* discrete state assignment parameters, where *K* is the number of class averages. Ab initio reconstructions(s) are calculated and used to refine *only* in-plane parameters by projection matching, leading to progressively improved class homogeneity.

RESULTS

Benchmarking on Simulated Data with Realistic Noise

Initial testing was carried out on a recently published heterogeneous phantom data set generated with a realistic noise model (structural noise, shot noise, and detector noise) to give a final signal-to-noise ratio of 0.05 (Baxter et al., 2009). The data are composed of 10,000 projections of two density maps of the 70S *E. coli* ribosome in two different states: one bound to elongation factor G (EF-G) with a single tRNA in the hybrid P/E position; and one lacking EF-G with tRNAs in the "classical" A/A, P/P, and E/E sites. The data were corrected for the sign shifts introduced by the CTF, centered, and rotationally aligned in 2D (Penczek et al., 1992). Principal component analysis was used in combination with hierarchical ascendant classification (Frank and van Heel, 1982; Lebart et al., 1984; van Heel and Frank, 1981) to group the data into 551 classes, each containing a minimum of 10 images. As expected, the first round of classification failed to completely resolve conformational heterogeneity due to inaccuracies in the in-plane alignment, giving 44% mixed class averages (Figure 2D) where a class was considered mixed if less than 75% of the particles were of the same state. After the first iteration, the state assignment algorithm had, rather than identifying true conformational state groups, separated three groups, containing about 10% of the data each, from the main body of correctly aligned, and better resolved averages. These groups corresponded to heterogeneous class averages (see Table 1). The reconstruction of one sparsely populated, and low-scoring group is shown in Figure 2C (gray). The state assignment algorithm discriminates poorly aligned groups of averages from well-aligned sets by assigning them to a state of their own. This useful property improves the quality of the initial reconstruction. In the first step, the main source of heterogeneity was primarily due to difference in quality of the class averages, and all averages of sufficient quality were accurately aligned and

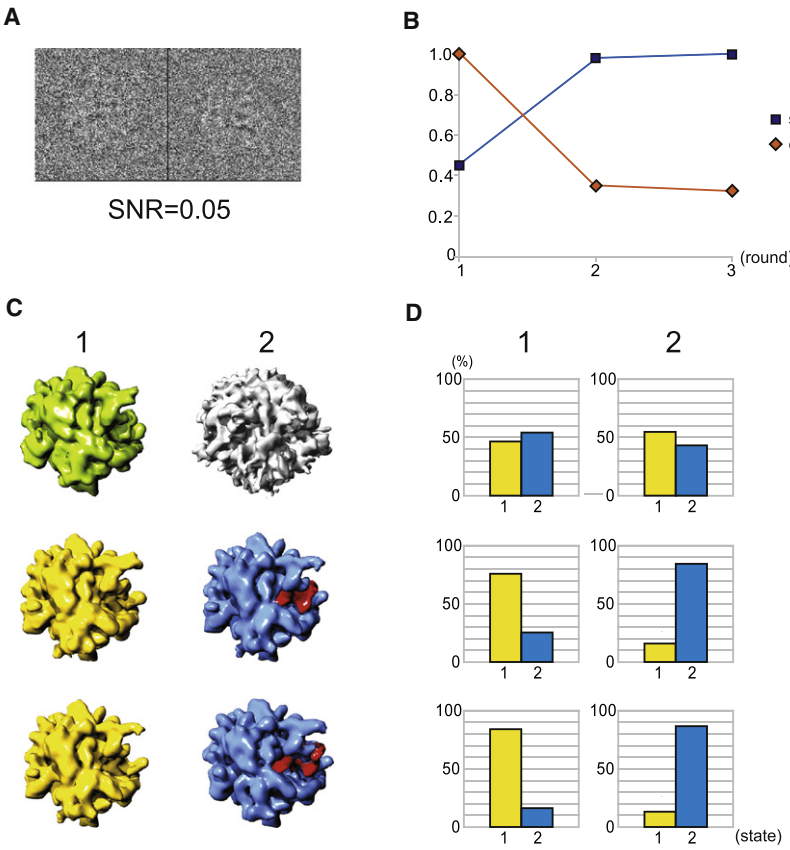

Figure 2. Benchmarking on Simulated Data with Realistic Noise

(A) Typical raw images processed only to correct for the sign shifts introduced by the CTF.

(B) Convergence plot with relative score values and classification errors indicated. The classification error is defined as the fraction of mixed class averages, where a class is considered mixed if less than 75% of the particles are of the same state.

(C) Reconstructions calculated for each round, with the EF-G lacking state (1) colored yellow, the EF-G containing state (2) colored blue, and the EF-G density colored red. The first reconstruction (green) is mixed due to in-plane alignment errors. In the first round, the conformational state assignment algorithm, rather than identifying true conformational state groups, separates lowly populated groups of very heterogeneous class averages from the main body of correctly aligned and better resolved averages. The high-quality state group is readily distinguished from the artificial ones by examining the score values. In following rounds classification and assignment separates the two conformational states.

(D) Percentage of correctly assigned individual particle images to each state.

grouped together to form a mixed first reconstruction (Figure 2C, green). This reconstruction was used as a template for in-plane alignment by projection matching. By the second iteration, the quality of the classification had improved significantly, resulting in only 17% mixed classes. EF-G bound class averages were now clearly separated from those lacking EF-G (Figures 2C and 2D). After the third round, we observed no significant improvement in quality of classification or score (Figure 2B), indicating that the process had reached convergence. The templates were used for refinement on individual images by competitive EvolAlign (Elmlund and Elmlund, 2009), resulting in two reconstructions (Figure S2) resolved to 15 Å according to the 0.5 Fourier Shell Correlation (FSC) criterion. The final distribution of particle images (5028 and 4972 images assigned to the EF-G bound and EF-G lacking state, respectively) was in perfect agreement with theory (99.7% correctly assigned images).

Table 1. First Iteration Statistics for Reconstruction 1–4

Reconstruction	Nr of Averages	Nr of Particles	Mixed Averages (%)
1	114	1392	38
2	54	782	41
3	51	1242	45
4	332	5849	25

Statistics include the number of class averages, the number of particles and the percentage of mixed class averages, where a class is considered mixed if less than 75% of the particles are of the same state.

Reconstruction of the *E. coli* 70S Ribosome Bound to Elongation Factor G

Next, the method was applied to a publicly available experimental cryo-EM data set containing 10,000 images of 70S *E. coli* ribosomes bound to EF-G (data available at http://www.ebi.ac.uk/msd/emdb/singleParticledir/SPIDER_FRANK_data/). This data set has become a popular test set for methods dealing with heterogeneity in cryo-EM (Elad et al., 2008; Scheres et al., 2007; Shatsky et al., 2010). As determined by supervised classification (Gao et al., 2004), the population was reported to be composed of an equal mixture of the two ribosome states described above. Unfortunately, the templates used for supervised classification are not provided with the data. The data were processed as described above and convergence was reached after four iterations. As expected, the final two reconstructions showed EF-G bound and EF-G lacking 70S ribosomes (Figure 3). In contrast to what has been reported for supervised classification (Gao et al., 2004) and ML-classification (Scheres et al., 2007) on larger but overlapping data sets, our EF-G lacking ribosome did not show any signs of tRNA density in the E/E position. Moreover, the L1 protuberance had swung away from the part of the intersubunit space that is overlapping with the E-site region. The negative correlation between the presence of the E-site tRNA and the L1 protuberance in the outward position is a well-known behavior that has been previously observed for the very same data (Elad et al., 2008) and for a similar EF-G bound ribosome data set (Penczek et al., 2006). Our final distribution of particle images (6703 and 3297 images assigned to the

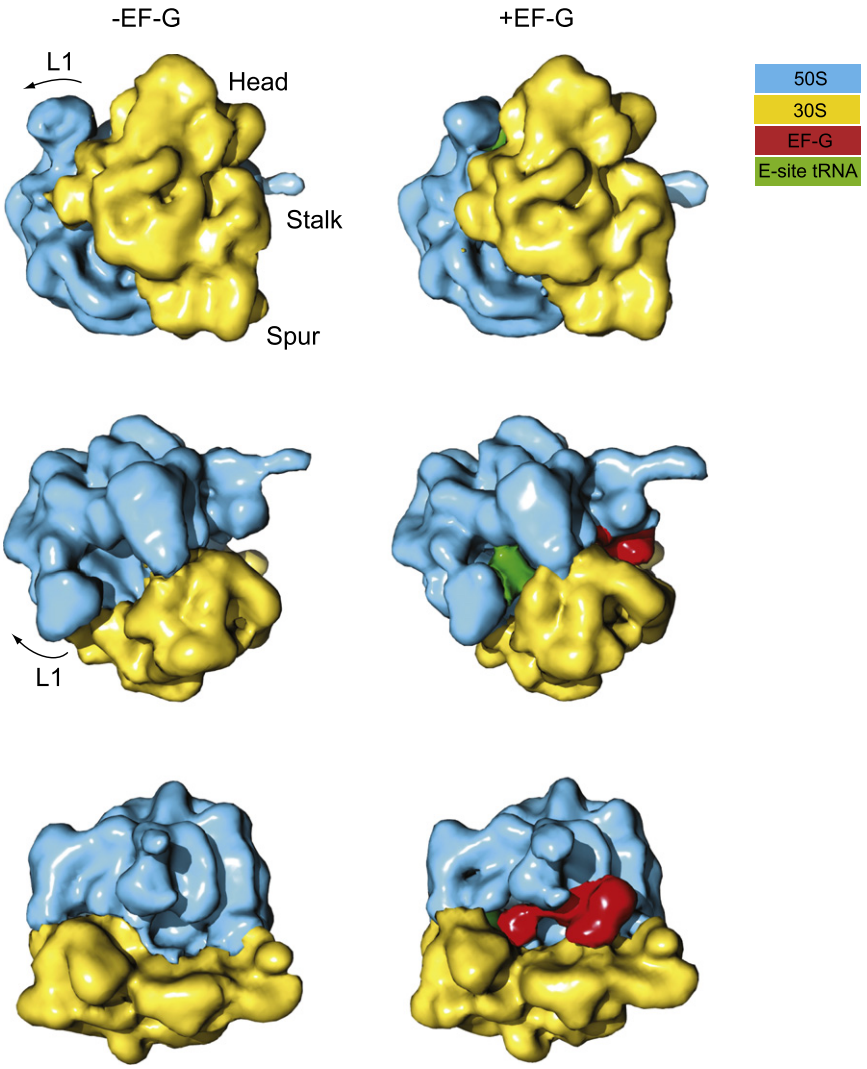

Figure 3. Ab Initio Cryo-EM Reconstructions of EF-G Bound and EF-G Lacking 70S Ribosomes

Well-known quaternary structure regions are indicated. The reconstructions differ primarily due to binding and release of EF-G. The EF-G associated state also shows the L1 protuberance in proximity of the E-site tRNA. Release of EF-G is coupled to a swinging motion of the L1 protuberance toward the solvent and dissociation of the E-site tRNA from the ribosome.

EF-G bound and EF-G lacking state, respectively) was in good agreement with what has been reported for the same data set (Elad et al., 2008).

Reconstruction of *Saccharomyces cerevisiae* RNA Polymerase II

Finally, the method was applied to 32771 EM images of individual *Saccharomyces cerevisiae* RNA polymerase II (pol II) particles in negative stain (see Experimental Procedures). Processing the data as described above, convergence was reached after six iterations. The two largest subunits of pol II, Rpb1, and Rpb2 form a positively charged "cleft," over which a massive "clamp" domain is swinging (Cramer et al., 2001). Subunits Rpb4 and Rpb7 form a dissociable heterodimer that binds to the base of the "clamp" near the mRNA exit site (Armache et al., 2005; Bushnell and Kornberg, 2003; Edwards et al., 1991). The final pol II reconstructions showed the 12-subunit enzyme in two distinctly different conformations, resolved to ~25 A (Figure 4). The position of the "clamp"-Rpb4/7 module within conformation 1 corresponds well to the crystal structure (Armache et al., 2003) (PDB 1WCM), whereas conformation 2 shows the enzyme in a "clamp-open" state, a previously unrecognized state for 12-subunit yeast pol II. To validate our interpretation, we modeled the clamp-open state of the 12-subunit crystal structure by docking the clamp-Rpb4/7 module from the "clamp-closed" 12-subunit structure to the clamp-open 10-subunit structure (Cramer et al., 2001) (PDB 15IO), revealing a 32.5 A movement of the clamp-Rpb4/7 module as measured between the Rpb4 subunits (Figure 4C). This movement coincided with the observed reconstructions (Figures 4A and 4B). The existence of a clamp-open 12-subunit yeast pol II structure is supported by a cryo-EM study of the human enzyme (Kostek et al., 2006) and allow a new interpretation of recent cross-linking experiments on yeast pol II (see Figure S3) (Chen et al., 2010).

DISCUSSION

In an attempt to reduce the model bias and increase the convergence radius of existing ab initio 3D reconstruction methods used in single-particle cryo-EM, this study addresses the problem of simultaneous reference-free 3D alignment and conformational state assignment of many thousands of single-particle projections. The method does not assume that a 3D alignment has been established beforehand by template-based alignment. At a first glance, our approach may seem convoluted. Why, for example, do we not use the template-based alignment that is done each round for 3D orientation assignment, but instead redo the time consuming classifications and RAD/GRASP/DE optimizations, over and over again. The answer is that we want to avoid the model bias that is inherent to all template-based alignment techniques. The effects of model bias or bias due to conformational heterogeneity when assuming a single underlying volume is not well known and detailed characterization of these phenomena will likely require a larger body of structural data at atomic resolution than that which is available today. Therefore, we have developed a technique that attempts to avoid using template-based alignment as far as possible. The inaccuracies of the first reference-free 2D alignment have to be accounted for (Penczek

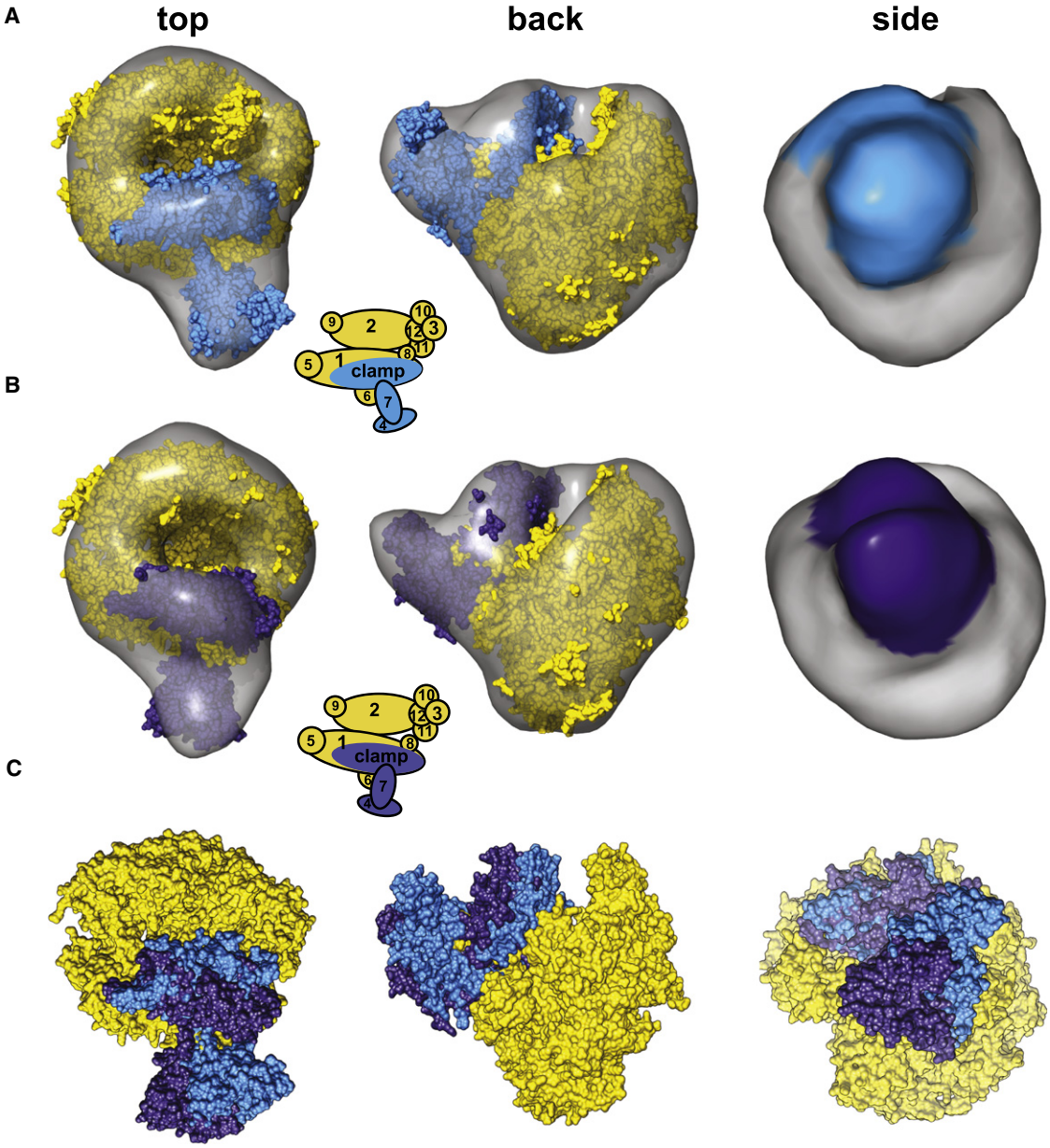

Figure 4. Ab Initio Negative Stain Reconstructions of Two RNA Polymerase II Conformations

Coloring key is indicated by the insets between the top and back views.

(A) 3D EM density map of pol II in "clamp-closed" state, with the 12 subunit clamp-closed X-ray structure docked (PDB 1WCM). The two models shown in (C) were fitted to the density, and the correlation coefficient calculated between the better fitting atomic model and the EM map was 0.7097, compared with 0.6218 for the worse fitting model.

(B) 3D EM density map of pol II in "clamp-open" state, with the modeled 12 subunit clamp-open structure docked. The correlation coefficient calculated between the better fitting atomic model and the EM map was 0.7545, compared with 0.6722 for the worse fitting model.

(C) Superposition of the clamp-closed and clamp-open structures showing the relative movement of the clamp-Rpb4/7 module.

et al., 1992), to force classification into resolving conformational heterogeneity, and for that we use projection matching onto the ab initio reconstructions. However, classification, orientation determination, and state assignment are all done independently of any template, which significantly reduces the model bias.

Heterogeneity of the kind introduced by partial factor occupancy is intensely studied in the ribosome field, and separation of states is often achieved by template matching, using one "empty" ribosome template, lacking the associated factor, and one template generated from the heterogeneous data (Gao et al., 2004; Valle et al., 2002). This method of "supervised classification" has been applied to the ribosome, but its general applicability is limited since it requires knowledge about the sample heterogeneity. To overcome this requirement, clusters

of images with similar conformational state may be identified by using protocols for focused classification (Elad et al., 2008; Fu et al., 2007; Penczek et al., 2006). The presence or absence of a ligand may be a relatively simple task to identify using these approaches, but it is not clear to us how state assignment will be achieved when conformational mobility affects several parts of a structure. Variance analysis may be used to guide focused classification and it has been shown to be a powerful tool on its own for characterization of structural variability (Zhang et al., 2008). Methods that do not depend on localization of specific structural regions include maximum likelihood (ML)-based approaches and methods based on common lines. Ab initio structure determination of coexisting conformational states via ML-classification has so far been achieved only for particles with icosahedral (60-fold) symmetry (Lee et al., 2007; Yin et al., 2003). However, by using available 3D reference volumes to provide a first 3D alignment, ML-classification has been used to resolve different states of the asymmetric ribosome (Scheres et al., 2007).

Our iterative data refinement scheme is based on calculation of common line correlations in Fourier space, and it bears similarity to the scheme proposed by Schatzky et al. (Shatsky et al., 2009, 2010). The major difference is that our approach combines reference-free 3D orientation determination with conformational state assignment and thereby circumvents the need for a priori structural data of asymmetric molecules. We confirm, by successful ab initio reconstruction from heterogeneous EM data sets of asymmetric molecules with known structures, that our proposed technique may be applicable to a wide range of single molecules coexisting in different functional states, be it macromolecular assemblies, nanoparticles, or colloids.

EXPERIMENTAL PROCEDURES

Purification of Yeast RNA Polymerase II

Pol II was purified from a strain with a TAP tag (Puig et al., 2001) integrated at the 3′ end of the Rpb3 gene (Wang et al., 2006). Cells were grown to late log/early stationary phase (OD600nm 9-10) in YPD media. After weighing wet cell pellet, cells were resuspended in 1 ml/g cells of 200 mM HEPES (pH 7.6), 10% glycerol, 2 mM EDTA, and 2 mM DTT with the protease inhibitors PMSF, benzamidine, leupeptin, and pepstatin. The slurry was adjusted to 20 mM ammonium sulfate with a saturated solution. Three hundred milliliters of slurry (~140 g of cells) and 225 ml glassbeads were used in one bead beater (BioSpec) for lysis. After cycling beater beader on 30 s/off 90 s for 45–50 min, the extract was debeaded by filtering through nylon mesh and centrifuged at 8K in a JA-14 rotor for 20 min. Bunk nucleic acids were removed by polyethyleneimine (PEI) precipitation by adjusting ammonium sulfate concentration to 200 mM and slowly adding buffered PEI to 0.3%. After 15 min stirring, the extract was centrifuged for 45 seconds at 14K in JA-14 rotor. The resulted clarified extract was precipitated with ammonium sulfate by adding an equal volume of saturated ammonium sulfate solution and then stirred for 30 min. The ammonium sulfate precipitate was collected by centrifuging 30 min at 14K in the JA-14 and decanting off supernatant. After resuspending the resultant pellet in 60 ml (25 mM HEPES [pH 7.6], 5% glycerol, 1 mM EDTA, and 2 mM DTT), the soluble extract was pumped through a 2 ml IgG fast flow Sepharose column at 20 ml/hr to capture the protein A-tagged pol II. The column was washed with 50 ml of the previous buffer plus 500 mM ammonium sulfate followed by 10 ml of 25 mM Tris (pH 7.8), 5% glycerol, 50 mM ammonium sulfate, 1 mM EDTA, and 2 mM DTT. The TAP tag was cleaved overnight (12–16 hr) by running 3 ml of the previous buffer containing 15 μg/ml TEV protease through the column at 4°C and capping for overnight cleavage. Pooled protein containing fractions were further purified by FPLC on an anion exchange column to yield 12-subunit RNA polymerase at homogeneity as assessed by SDS-PAGE.

Negative Stain Specimen Preparation, Electron Microscopy, and Image Processing

Four microliters of pol II solution, diluted 300 times in buffer solution without glycerol, was applied to freshly glow-discharged, carbon-coated 400-mesh Cu grids (Electron Microscopy Sciences) that were incubated for 30 s before removal of excess solvent. The specimens were stained by application of 3 μl 2% uranyl formate solution. Excess staining solution was immediately removed by blotting with a filter paper, and the staining procedure was repeated three times. CCD images were recorded under low-dose conditions ($15e^-/A^2$) at 0.1–0.3 μm underfocus, using a FEI Tecnai F20 transmission electron microscope equipped with a field-emission gun, and operating at an acceleration voltage of 200 kV. Data were collected at 25k magnification, giving a sampling distance of 4.62 A at the specimen level. The CCD images were low-pass filtered to the first zero crossing of the CTF before automatic windowing of individual particles in Boxer (Ludtke et al., 1999). USCF Chimera was used for automatic real-space rigid-body docking and visualization (Pettersen et al., 2004). Spider was used for calculation of Fourier transforms, CTF-correction, classification, and projection matching (Frank et al., 1996).

SUPPLEMENTAL INFORMATION

Supplemental Information includes Supplemental Experimental Procedures and three figures and can be found with this article online at doi:10.1016/j.str.2010.06.001.

ACKNOWLEDGMENTS

We thank Karl-Magnus Larsson, Philip Robinson, and David Bushnell for critical reading of the manuscript and helpful suggestions; the LUNARC center for distributed computing at Lund University, Sweden. H.E. was supported by grants from the Swedish Research Council and the Wenner-Gren foundation. D.E. was supported by the Marie-Curie Early-Stage Training grant (ECfp7).

Received: February 5, 2010
Revised: June 6, 2010
Accepted: June 7, 2010
Published: July 13, 2010

REFERENCES

Adrian, M., Dubochet, J., Lepault, J., and McDowall, A.W. (1984). Cryo-electron microscopy of viruses. Nature *308*, 32–36.

Armache, K.J., Kettenberger, H., and Cramer, P. (2003). Architecture of initiation-competent 12-subunit RNA polymerase II. Proc. Natl. Acad. Sci. USA *100*, 6964–6968.

Armache, K.J., Mitterweger, S., Meinhart, A., and Cramer, P. (2005). Structures of complete RNA polymerase II and its subcomplex, Rpb4/7. J. Biol. Chem. *280*, 7131–7134.

Baxter, W.T., Grassucci, R.A., Gao, H.X., and Frank, J. (2009). Determination of signal-to-noise ratios and spectral SNRs in cryo-EM low-dose imaging of molecules. J. Struct. Biol. *166*, 126–132.

Bracewell, R.N. (1956). Strip integration in radio astronomy. Aust. J. Phys. *9*, 198–217.

Brink, J., Ludtke, S.J., Kong, Y.F., Wakil, S.J., Ma, J.P., and Chiu, W. (2004). Experimental verification of conformational variation of human fatty acid synthase as predicted by normal mode analysis. Structure *12*, 185–191.

Bushnell, D.A., and Kornberg, R.D. (2003). Complete, 12-subunit RNA polymerase II at 4.1-angstrom resolution: Implications for the initiation of transcription. Proc. Natl. Acad. Sci. USA *100*, 6969–6973.

Chen, Z.A., Jawhari, A., Fischer, L., Buchen, C., Tahir, S., Kamenski, T., Rasmussen, M., Lariviere, L., Bukowski-Wills, J.C., Nilges, M., et al. (2010). Architecture of the RNA polymerase II-TFIIF complex revealed by cross-linking and mass spectrometry. EMBO J *29*, 717–726.

Cramer, P., Bushnell, D.A., and Kornberg, R.D. (2001). Structural basis of transcription: RNA polymerase II at 2.8 angstrom ngstrom resolution. Science *292*, 1863–1876.

Crowther, R.A., Amos, L.A., Finch, J.T., Derosier, D.J., and Klug, A. (1970a). Three dimensional reconstructions of spherical viruses by fourier synthesis from electron micrographs. Nature *226*, 421–425.

Crowther, R.A., Derosier, D.J., and Klug, A. (1970b). Reconstruction of three-dimensional structure from projections and its application to electron microscopy. Proc. R. Soc. Lond. A *317*, 319–340.

DeRosier, D.J., and Klug, A. (1968). Reconstruction of three-dimensional structures from electron micrographs. Nature *217*, 130–134.

Edwards, A.M., Kane, C.M., Young, R.A., and Kornberg, R.D. (1991). Two dissociable subunits of yeast rna polymerase-II stimulate the initiation of transcription at a promoter in vitro. J. Biol. Chem. *266*, 71–75.

Elad, N., Clare, D.K., Salbil, H.R., and Orlova, E.V. (2008). Detection and separation of heterogeneity in molecular complexes by statistical analysis of their two-dimensional projections. J. Struct. Biol. *162*, 108–120.

Elmlund, D., and Elmlund, H. (2009). High-resolution single-particle orientation refinement based on spectrally self-adapting common lines. J. Struct. Biol. *167*, 83–94.

Elmlund, H., Lundqvist, J., Al-Karadaghi, S., Hansson, M., Hebert, H., and Lindahl, M. (2008). A new cryo-EM single-particle *ab initio* reconstruction method visualizes secondary structure elements in an ATP-fuelled AAA+ motor. J. Mol. Biol. *375*, 934–947.

Elmlund, H., Baraznenok, V., Linder, T., Szilagyi, Z., Rofugaran, R., Hofer, A., Hebert, H., Lindahl, M., and Gustafsson, C.M. (2009). Cryo-EM reveals promoter DNA binding and conformational flexibility of the general transcription factor TFIID. Structure *17*, 1442–1452.

Feo, T.A., and Resende, M.G.C. (1995). Greedy randomized adaptive search procedures. J. Global Optimization *6*, 109–133.

Frank, J. (2006). Three-Dimensional Electron Microscopy of Macromolecular Assemblies, Volume 2 (New York: Oxford University Press).

Frank, J., Radermacher, M., Penczek, P., Zhu, J., Li, Y.H., Ladjadj, M., and Leith, A. (1996). SPIDER and WEB: processing and visualization of images in 3D electron microscopy and related fields. J. Struct. Biol. *116*, 190–199.

Frank, J., and van Heel, M. (1982). Correspondence-analysis of aligned images of biological particles. J. Mol. Biol. *161*, 134–137.

Fu, J., Gao, H.X., and Frank, J. (2007). Unsupervised classification of single particles by cluster tracking in multi-dimensional space. J. Struct. Biol. *157*, 226–239.

Gao, H., Valle, M., Ehrenberg, M., and Frank, J. (2004). Dynamics of EF-G interaction with the ribosome explored by classification of a heterogeneous cryo-EM dataset. J. Struct. Biol. *147*, 283–290.

Hall, R.J., Siridechadilok, B., and Nogales, E. (2007). Cross-correlation of common lines: a novel approach for single-particle reconstruction of a structure containing a flexible domain. J. Struct. Biol. *159*, 474–482.

Harauz, G., and van Heel, M. (1986). Exact filters for general geometry 3-dimensional reconstruction. Optik (Stuttg.) *73*, 146–156.

Herman, G.T., and Kalinowski, M. (2007). Clasification of heterogeneous electron microscopic projections into homogeneous subsets. Ultramicroscopy *108*, 327–338.

Hoppe, W., Schramm, H.J., Sturm, M., Hunsmann, N., and Gassmann, J. (1976). 3-dimensional electron-microscopy of individual biological objects. Z. Naturforsch. C *31*, 645–655.

Kirkpatrick, S., Gelatt, C.D., and Vecchi, M.P. (1983). Optimization by simulated annealing. Science *220*, 671–680.

Knapek, E., and Dubochet, J. (1980). Beam damage to organic material is considerably reduced in cryo-electron microscopy. J. Mol. Biol. *141*, 147–161.

Kostek, S.A., Grob, P., De Carlo, S., Lipscomb, J.S., Garczarek, F., and Nogales, E. (2006). Molecular architecture and conformational flexibility of human RNA polymerase II. Structure *14*, 1691–1700.

Lebart, L., Morineau, A., and Warnick, K.M. (1984). Multivariate descriptive statistical analysis (New York: Wiley).

Lee, J., Doerschuk, P.C., and Johnson, J.E. (2007). Exact reduced-complexity maximum likelihood reconstruction of multiple 3-d objects from unlabeled unoriented 2-d projections and electron microscopy of viruses. IEEE Trans. Image Process *16*, 2865–2878.

Lindahl, M. (2001). Strul - a method for 3D alignment of single-particle projections based on common line correlation in Fourier space. Ultramicroscopy *87*, 165–175.

Ludtke, S.J., Baldwin, P.R., and Chiu, W. (1999). EMAN: semiautomated software for high-resolution single-particle reconstructions. J. Struct. Biol. *128*, 82–97.

Lundqvist, J., Elmlund, H., Wulff, R.P., Berglund, L., Elmlund, D., Emanuelsson, C., Hebert, H., Willows, R.D., Hansson, M., Lindahl, M., and Al-Karadaghi, S. (2010). ATP-induced conformational dynamics in the AAA plus motor unit of magnesium chelatase. Structure *18*, 354–365.

Metropolis, N., and Ulam, S. (1949). The Monte Carlo Method. J. Am. Stat. Assoc. *44*, 335–341.

Metropolis, N., Rosenbluth, A.W., Rosenbluth, M.N., Teller, A.H., and Teller, E. (1953). Equation of state calculations by fast computing machines. J. Chem. Phys. *21*, 1087–1092.

Ogura, T., and Sato, C. (2006). A fully automatic 3D reconstruction method using simulated annealing enables accurate posterioric angular assignment of protein projections. J. Struct. Biol. *156*, 371–386.

Penczek, P., Radermacher, M., and Frank, J. (1992). 3-dimensional reconstruction of single particles embedded in ice. Ultramicroscopy *40*, 33–53.

Penczek, P.A., Grassucci, R.A., and Frank, J. (1994). The ribosome at improved resolution—new techniques for merging and orientation refinement in 3D cryoelectron microscopy of biological particles. Ultramicroscopy *53*, 251–270.

Penczek, P.A., Zhu, J., and Frank, J. (1996). A common-lines based method for determining orientations for N>3 particle projections simultaneously. Ultramicroscopy *63*, 205–218.

Penczek, P.A., Frank, J., and Spahn, C.M.T. (2006). A method of focused classification, based on the bootstrap 3D variance analysis, and its application to EF-G-dependent translocation. J. Struct. Biol. *154*, 184–194.

Pettersen, E.F., Goddard, T.D., Huang, C.C., Couch, G.S., Greenblatt, D.M., Meng, E.C., and Ferrin, T.E. (2004). UCSF chimera—a visualization system for exploratory research and analysis. J. Comput. Chem. *25*, 1605–1612.

Puig, O., Caspary, F., Rigaut, G., Rutz, B., Bouveret, E., Bragado-Nilsson, E., Wilm, M., and Seraphin, B. (2001). The tandem affinity purification (TAP) method: a general procedure of protein complex purification. Methods *24*, 218–229.

Radermacher, M. (1992). Weighted Back-Projection Methods (New York: Plenum Press).

Saff, E.B., and Kuijlaars, A.B.J. (1997). Distributing many points on a sphere. Mathematical Intelligencer *19*, 5–11.

Scheres, S.H.W., Gao, H.X., Valle, M., Herman, G.T., Eggermont, P.P.B., Frank, J., and Carazo, J.M. (2007). Disentangling conformational states of macromolecules in 3D-EM through likelihood optimization. Nat. Methods *4*, 27–29.

Shatsky, M., Hall, R.J., Brenner, S.E., and Glaeser, R.M. (2009). A method for the alignment of heterogeneous macromolecules from electron microscopy. J. Struct. Biol. *166*, 67–78.

Shatsky, M., Hall, R.J., Nogales, E., Malik, J., and Brenner, S. (2010). Automated multi-model reconstruction from single-particle electron microscopy data. J. Struct. Biol. *170*, 98–108.

Singer, A., Coifman, R.R., Sigworth, F.J., Chester, D.W., and Shkolnisky, Y. (2009). Detecting consistent common lines in cryo-EM by voting. J. Struct. Biol. *169*, 312–322.

Valle, M., Sengupta, J., Swami, N.K., Grassucci, R.A., Burkhardt, N., Nierhaus, K.H., Agrawal, R.K., and Frank, J. (2002). Cryo-EM reveals an active role for aminoacyl-tRNA in the accommodation process. EMBO J. *21*, 3557–3567.

van Heel, M. (1987). Angular reconstitution—a posteriori assignment of projection directions for 3-D reconstruction. Ultramicroscopy *21*, 111–123.

van Heel, M., and Frank, J. (1981). Use of multivariate statistics in analyzing the images of biological macromolecules. Ultramicroscopy *6*, 187–194.

van Heel, M., Gowen, B., Matadeen, R., Orlova, E.V., Finn, R., Pape, T., Cohen, D., Stark, H., Schmidt, R., Schatz, M., and Patwardhan, A. (2000). Single-particle electron cryo-microscopy: towards atomic resolution. Q. Rev. Biophys. *33*, 307–369.

Wang, D., Bushnell, D.A., Westover, K.D., Kaplan, C.D., and Kornberg, R.D. (2006). Structural basis of transcription: Role of the trigger loop in substrate specificity and catalysis. Cell *127*, 941–954.

Yin, Z., Zheng, Y., Doerschuk, P.C., Natarajan, P., and Johnson, J.E. (2003). A statistical approach to computer processing of cryo-electron microscope images: virion classification and 3-D reconstruction. J. Struct. Biol. *144*, 24–50.

Zhang, W., Kirnmel, M., Spahn, C.M.T., and Penczek, P.A. (2008). Heterogeneity of Large Macromolecular Complexes Revealed by 3D Cryo-EM Variance Analysis. Structure *16*, 1770–1776.

